Proceedings of the
46th IEEE Conference on Decision and Control
New Orleans, LA, USA, Dec. 12-14, 2007

Payoff Based Dynamics for Multi-Player Weakly Acyclic Games

Jason R. Marden, H. Peyton Young, Gürdal Arslan and Jeff S. Shamma

Abstract— **We consider repeated multi-player games in which players repeatedly and simultaneously choose strategies from a finite set of available strategies according to some strategy adjustment process. We focus on the specific class of weakly acyclic games, which is particularly relevant for multi-agent cooperative control problems. A strategy adjustment process determines how players select their strategies at any stage as a function of the information gathered over previous stages. Of particular interest are "payoff based" processes, in which at any stage, players only know their own actions and (noise corrupted) payoffs from previous stages. In particular, players do not know the actions taken by other players and do not know the structural form of payoff functions. We introduce three different payoff based processes for increasingly general scenarios and prove that after a sufficiently large number of stages, player actions constitute a Nash equilibrium at any stage with arbitrarily high probability. We also show how to modify player utility functions through tolls and incentives in so-called congestion games, a special class of weakly acyclic games, to guarantee that a centralized objective can be realized as a Nash equilibrium. We illustrate the methods with a simulation of distributed routing over a network.**

I. Introduction

The objective in distributed cooperative control for multi-agent systems is to enable a collection of "self-interested" agents to achieve a desirable "collective" objective. There are two overriding challenges to achieving this objective. The first is complexity: finding an optimal solution by a centralized algorithm may be prohibitively difficult when there are large numbers of interacting agents. This motivates the use of adaptive methods that enable agents to "self organize" into suitable, if not optimal, collective solutions.

The second challenge is limited information. Agents may have limited knowledge about the status of other agents, except perhaps for a small subset of "neighboring" agents. An example is collective motion control for mobile sensor platforms (e.g., [2]). In these problems, mobile sensors seek to position themselves to achieve various collective objectives such as rendezvous or area coverage. Sensors can communicate with neighboring sensors, but otherwise do not have global knowledge of the domain of operation or the status and locations of non-neighboring sensors.

A typical assumption is that agents are endowed with a reward or utility function that depends on their own strategies and the strategies of other agents. In motion coordination problems, for example, an agent's utility function typically depends on its position relative to other agents or environmental targets, and knowledge of this function guides local motion adjustments.

In other situations, agents may know nothing about the structure of their utility functions, and how their own utility depends on the actions of other agents (whether local or far away). In this case the only thing they can do is observe rewards based on experience and "optimize" on a trial and error basis. The situation is further complicated because all agents are trying simultaneously to optimize their own strategies. Therefore, even in the absence of noise, an agent trying the same strategy twice may see different results because of the non-stationary nature of the strategies of other agents.

There are several examples of multi-agent systems that illustrate this situation. In distributed routing for ad hoc data networks (e.g., [3]), routing nodes seek to route packets to neighboring nodes based on packet destinations without knowledge of the overall network structure. The objective is to minimize the delay of packets to their destinations. This delay must be realized through trial and error, since the functional dependence of delay on routing strategies is not known. A similar problem is automotive traffic routing, in which drivers seek to minimize the congestion experienced to get to a desired destination. Drivers can experience the congestion on selected routes as a function of the routes selected by other drivers, but drivers do not know the structure of the congestion function. Finally, in a multi-agent approach to designing manufacturing systems (e.g., [4]), it may not be known in advance how performance measures (such as throughput) depend on manufacturing policy. Rather performance can only be measured once a policy is implemented.

Our interest in this paper is to develop algorithms that enable coordination in multi-agent systems for precisely this "payoff based" scenario, in which agents only have access to (possibly noisy) measurements of the rewards received through repeated interactions with other agents. We adopt the framework of "learning in games" (see [5], [6], [7], [8] for an extensive overview). Unlike most of the learning rules in this literature, which assume that agents adjust their behavior based on the observed behavior of other agents, we shall assume that agents know only their own past actions and the payoffs that resulted. It is far from obvious that Nash equilibrium can be achieved under such a restriction, but in

Research supported by NSF grant #ECS-0501394, ARO grant #W911NF–04–1–0316, and AFOSR grant #FA9550-05-1-0239. The journal version of this papers appears in [1].

J. R. Marden is with the Social and Information Sciences Laboratory, California Institute of Technology, 1200 E. California Blvd., MC 136-93, Pasadena, CA 91125, marden@caltech.edu.

H. P. Young is with the Department of Economics, Johns Hopkins University, 440 Mergenthaler Hall 3400, N. Charles Street Baltimore, MD 21218, pyoung@jhu.edu. H. P. Young is also with the Department of Economics, University of Oxford and the Center on Social and Economic Dynamics at the Brookings Institution in Washington.

G. Arslan is with the Department of Electrical Engineering, University of Hawaii at Manoa, 440 Holmes Hall, 2540 Dole Street, Honolulu, HI 96822, gurdal@hawaii.edu.

J. S. Shamma is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, 777 Atlantic Dr NW, Atlanta, GA 30332-0250, shamma@gatech.edu.

1-4244-1498-9/07/$25.00 ©2007 IEEE.

fact it has recently been shown that such "payoff based" learning rules can be constructed that work in any game [9], [10].

In this paper we show that there are simpler and more intuitive adjustment rules that achieve this objective for a large class of multi-player games known as "weakly acyclic" games. This class captures many problems of interest in cooperative control [11], [12]. It includes the very special case of "identical interest" games, where each agent receives the same reward. However, weakly acyclic games (and the related concept of potential games) capture other scenarios such as congestion games [13] and similar problems such as distributed routing in networks, weapon target assignment, consensus, and area coverage. See [14], [15] and referenced therein for a discussion of a learning in games approach to cooperative control problems, but under less stringent assumptions on informational constraints considered in this paper.

For many multi-agent problems, operation at a pure Nash equilibrium may reflect optimization of a collective objective.[1] We will derive payoff based dynamics that guarantee asymptotically that agent strategies will constitute a pure Nash equilibrium with arbitrarily high probability. It need not always be the case that at least one Nash equilibrium optimizes a collective objective. Motivated by this consideration, we also discuss the introduction of incentives or tolls in a player's payoff function to assure that there is at least one Nash equilibrium that optimizes a collective objective. Even in this case, however, there may still be suboptimal Nash equilibria.

The remainder of this paper is organized as follows. Section 2 provides background on finite strategic-form games and repeated games. This is followed by three types of payoff based dynamics in Section 3 for increasingly general problems. Section 3.1 presents "Safe Experimentation Dynamics" which is restricted to identical interest games. Section 3.2 presents "Simple Experimentation Dynamics" for the more general class of weakly acyclic games but with noise free payoff measurements. Section 3.3 presents "Sample Experimentation Dynamics" for weakly acyclic games with noisy payoff measurements. Section 4 discusses how to introduce tolls and incentives in payoffs so that a Nash equilibrium optimizes a collective objective. Section 5 presents an illustrative example of a traffic congestion game. Finally, Section 6 contains some concluding remarks. An important analytical tool throughout is the method of resistance trees for perturbed Markov chains [18], which is reviewed in the appendix of [1]. We will omit many of the proofs for brevity. The complete proofs can be found in the journal version of this paper [1].

II. Background

In this section, we will present a brief background of the game theoretic concepts used in the paper. We refer the readers to [19], [7], [8] for a more comprehensive review.

A. Finite Strategic-Form Games

Consider a finite strategic-form game with n-player set $\mathcal{P} := \{\mathcal{P}_1, ..., \mathcal{P}_n\}$ where each player $\mathcal{P}_i \in \mathcal{P}$ has an action set Y_i and a utility function $U_i : Y \to \mathcal{R}$ where $Y = Y_1 \times ... \times Y_n$. We will sometimes use a single symbol, e.g., G, to represent the entire game, i.e., the player set, $\mathcal{P}$, action sets, Y_i, and utility functions U_i.

For an action profile $y = (y_1, y_2, ..., y_n) \in Y$, let y_{-i} denote the profile of player actions *other than* player $\mathcal{P}_i$, i.e., $y_{-i} = \{y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n\}$. With this notation, we will sometimes write a profile y of actions as (y_i, y_{-i}). Similarly, we may write $U_i(y)$ as $U_i(y_i, y_{-i})$.

An action profile $y^* \in Y$ is called a *pure Nash equilibrium* if for all players $\mathcal{P}_i \in \mathcal{P}$,

$$U_i(y_i^*, y_{-i}^*) = \max_{y_i \in Y_i} U_i(y_i, y_{-i}^*). \quad (1)$$

In this paper we will consider three classes of games: identical interest games, potential games, and weakly acyclic games. Each class of games has a connection to general cooperative control problems and multi-agent systems for which there is some global utility or potential function $\phi : Y \to \mathcal{R}$ that a global planner seeks to maximize [11].

1) Identical Interest Games: The most restrictive class of games that we will review in this paper is identical interest games. In such a game, the players' utility functions $\{U_i\}_{i=1}^n$ are chosen to be the same. That is, for some function $\phi : Y \to \mathcal{R}$, $U_i(y) = \phi(y)$, for every $\mathcal{P}_i \in \mathcal{P}$ and for every $y \in Y$. It is easy to verify that all identical interest games have at least one pure Nash equilibrium, namely any action profile y that maximizes $\phi(y)$.

2) Potential Games: A significant generalization of an identical interest game is a potential game. In a potential game, the change in a player's utility that results from a unilateral change in strategy equals the change in the global utility. Specifically, there is a function $\phi : Y \to \mathcal{R}$ such that for every player $\mathcal{P}_i \in \mathcal{P}$, for every $y_{-i} \in Y_{-i}$, and for every $y_i', y_i'' \in Y_i$,

$$U_i(y_i', y_{-i}) - U_i(y_i'', y_{-i}) = \phi(y_i', y_{-i}) - \phi(y_i'', y_{-i}).$$

When this condition is satisfied, the game is called a potential game with the potential function ϕ. It is easy to see that, in potential games, any action profile maximizing the potential function is a pure Nash equilibrium, hence every potential game possesses at least one such equilibrium.

3) Weakly Acyclic Games: Consider any finite game G with a set Y of action profiles. A *better reply path* is a sequence of action profiles $y^1, y^2, ..., y^L$ such that, for every $1 \leq \ell \leq L-1$, there is exactly one player $\mathcal{P}_{i_\ell}$ such that i) $y_{i_\ell}^\ell \neq y_{i_\ell}^{\ell+1}$, ii) $y_{-i_\ell}^\ell = y_{-i_\ell}^{\ell+1}$, and iii) $U_{i_\ell}(y^\ell) < U_{i_\ell}(y^{\ell+1})$. In other words, one player moves at a time, and each time a player moves he increases his own utility.

Suppose now that G is a potential game with potential function ϕ. Starting from an arbitrary action profile $y \in Y$, construct a better reply path $y = y^1, y^2, ..., y^L$ until it can no longer be extended. Note first that such a path cannot cycle back on itself, because ϕ is strictly increasing along the path. Since Y is finite, the path cannot be extended indefinitely. Hence, the last element in a maximal better reply path from any joint action, y, must be a Nash equilibrium of G.

This idea may be generalized as follows. The game G is *weakly acyclic* if for any $y \in Y$, there exists a better reply path starting at y and ending at some pure Nash equilibrium of G [7], [8]. Potential games are special cases of weakly acyclic games.

[1] Nonetheless, there are varied viewpoints on the role of Nash equilibrium as a solution concept for multi-agent systems. See [16] and [17].

B. Repeated Games

In a repeated game, at each time $t \in \{0, 1, 2, \dots\}$, each player $\mathcal{P}_i \in \mathcal{P}$ simultaneously chooses an action $y_i(t) \in Y_i$ and receives the utility $U_i(y(t))$ where $y(t) := (y_1(t), \dots, y_n(t))$. Each player $\mathcal{P}_i \in \mathcal{P}$ chooses his action $y_i(t)$ at time t according to a probability distribution $p_i(t)$, which we will refer to as the *strategy* of player $\mathcal{P}_i$ at time t. A player's strategy at time t can rely only on observations from times $\{0, 1, 2, ..., t-1\}$. Different learning algorithms are specified by both the assumptions on available information and the mechanism by which the strategies are updated as information is gathered. For example, if a player knows the functional form of his utility function and is capable of observing the actions of all other players at every time step, then the strategy adjustment mechanism of player $\mathcal{P}_i$ can be written in the general form

$$p_i(t) = F_i\big(y(0), ..., y(t-1); U_i\big).$$

An example of a learning algorithm, or strategy adjustment mechanism, of this form is the well known fictitious play [20]. For a detailed review of learning in games we direct the reader to [5], [7], [8], [21], [22], [23].

In this paper we deal with the issue of whether players can learn to play a pure Nash equilibrium through repeated interactions under the most restrictive observational conditions; players *only* have access to (i) the action they played and (ii) the utility (possibly noisy) they received. In this setting, the strategy adjustment mechanism of player $\mathcal{P}_i$ takes on the form

$$p_i(t) = F_i\big(\{y_i(k), U_i(y(k)) + \nu_i(k)\}_{k=0,1,\dots,t-1}\big),$$

where the $\nu_i(t)$ are zero mean independent and identically distributed (i.i.d.) random variables.

III. Payoff Based Learning Algorithms

In this section, we will introduce three simple payoff based learning algorithms. The first, called *Safe Experimentation*, guarantees convergence to a pure optimal Nash equilibrium in any identical interest game. Such an equilibrium is optimal because each player's utility is maximized. The second learning algorithm, called *Simple Experimentation*, guarantees convergence to a pure Nash equilibrium in any weakly acyclic game. The third learning algorithm, called *Sample Experimentation*, guarantees convergence to a pure Nash equilibrium in any weakly acyclic game even when utility measurements are corrupted with noise.

A. Safe Experimentation Dynamics for Identical Interest Games

Before introducing the learning dynamics, we introduce the following function. Let

$$U_i^{\max}(t) := \max_{0 \le \tau \le t-1} U_i(y(\tau))$$

be the maximum utility that player $\mathcal{P}_i$ has received up to time $t-1$.

We will now introduce the Safe Experimentation dynamics for identical interest games.

1) **Initialization:** At time $t = 0$, each player randomly selects and plays any action, $y_i(0)$. This action will be initially set as the player's *baseline action* at time $t = 1$ and is denoted by $y_i^b(1) = y_i(0)$.
2) **Action Selection:** At each subsequent time step, each player selects his baseline action with probability $(1-\epsilon)$ or experiments with a new random action with probability ϵ, i.e.:
   - $y_i(t) = y_i^b(t)$ with probability $(1-\epsilon)$
   - $y_i(t)$ is chosen randomly (uniformly) over Y_i with probability ϵ

   The variable ϵ will be referred to as the player's *exploration rate*.
3) **Baseline Strategy Update:** Each player compares the actual utility received, $U_i(y(t))$, with the maximum received utility $U_i^{\max}(t)$ and updates his baseline action as follows:

$$y_i^b(t+1) = \begin{cases} y_i(t), & U_i(y(t)) > U_i^{\max}(t); \\ y_i^b(t), & U_i(y(t)) \le U_i^{\max}(t). \end{cases}$$

   This step is performed whether or not Step 2 involved exploration.
4) Return to Step 2 and repeat.

The reason that this learning algorithm is called "Safe" Experimentation is that the utility evaluated at the baseline action, $U(y^b(t))$, is non-decreasing with respect to time.

Theorem 3.1: Let G be a finite n-player identical interest game in which all players use the Safe Experimentation dynamics. Given any probability $p < 1$, if the exploration rate $\epsilon > 0$ is sufficiently small, then for all sufficiently large times t, $y(t)$ is an optimal Nash equilibrium of G with at least probability p.

The proof is omitted for brevity. See [1].

B. Simple Experimentation Dynamics for Weakly Acyclic Games

We will now introduce the Simple Experimentation dynamics for weakly acyclic games. These dynamics will allow us to relax the assumption of identical interest games.

1) **Initialization:** At time $t = 0$, each player randomly selects and plays any action, $y_i(0)$. This action will be initially set as the player's *baseline action* at time 1, i.e., $y_i^b(1) = y_i(0)$. Likewise, the player's *baseline utility* at time 1 is initialized as $u_i^b(1) = U_i(y(0))$.
2) **Action Selection:** At each subsequent time step, each player selects his baseline action with probability $(1-\epsilon)$ or experiments with a new random action with probability ϵ.
   - $y_i(t) = y_i^b(t)$ with probability $(1-\epsilon)$
   - $y_i(t)$ is chosen randomly (uniformly) over Y_i with probability ϵ

   The variable ϵ will be referred to as the player's *exploration rate*. Whenever $y_i(t) \neq y_i^b(t)$, we will say that player $\mathcal{P}_i$ *experimented*.
3) **Baseline Action and Baseline Utility Update:** Each player compares the utility received, $U_i(y(t))$, with his baseline utility, $u_i^b(t)$, and updates his baseline action and utility as follows:
   - If player $\mathcal{P}_i$ *experimented* (i.e., $y_i(t) \neq y_i^b(t)$) and if $U_i(y(t)) > u_i^b(t)$ then
     - $y_i^b(t+1) = y_i(t)$,
     - $u_i^b(t+1) = U_i(y(t))$.
   - If player $\mathcal{P}_i$ *experimented* and if $U_i(y(t)) \le u_i^b(t)$ then
     - $y_i^b(t+1) = y_i^b(t)$,
     - $u_i^b(t+1) = u_i^b(t)$.
   - If player $\mathcal{P}_i$ *did not experiment* (i.e., $y_i(t) = y_i^b(t)$) then
     - $y_i^b(t+1) = y_i^b(t)$,
     - $u_i^b(t+1) = U_i(y(t))$.

4) Return to Step 2 and repeat.

As before, these dynamics require only utility measurements, and hence almost no information regarding the structure of the game.

Theorem 3.2: Let G be a finite n-player weakly acyclic game in which all players use the Simple Experimentation dynamics. Given any probability $p < 1$, if the exploration rate $\epsilon > 0$ is sufficiently small, then for all sufficiently large times t, $y(t)$ is a Nash equilibrium of G with at least probability p.

The proof is omitted for brevity. See [1].

C. Sample Experimentation Dynamics for Weakly Acyclic Games with Noisy Utility Measurements

In this section we will focus on developing payoff based dynamics for which the limiting behavior exhibits that of a pure Nash equilibrium with arbitrarily high probability in any finite weakly acyclic game *even in the presence of utility noise*. We will show that a variant of the so-called Regret Testing algorithm [9] accomplishes this objective for weakly acyclic games with noisy utility measurements.

We now introduce Sample Experimentation dynamics.

1) **Initialization:** At time $t = 0$, each player randomly selects and plays any action, $y_i(0) \in Y_i$. This action will be initially set as the player's *baseline action*, $y_i^b(1) = y_i(0)$.
2) **Exploration Phase:** After the baseline action is set, each player engages in an *exploration phase* over the next m periods. The length of the exploration phase need not be the same or synchronized for each player. For convenience, we will double index the time of the actions played as
$$\check{y}(t_1, t_2) = y(m\ t_1 + t_2)$$
where t_1 indexes the number of the exploration phase and t_2 indexes the actions played in that exploration phase. We will refer to t_1 as the *exploration phase time* and t_2 as the *exploration action time*. By construction, the exploration phase time and exploration action time satisfy $t_1 \geq 1$ and $m \geq t_2 \geq 1$. The baseline action will only be updated at the end of the exploration phase and will therefore only be indexed by the exploration phase time.
During the exploration phase, each player selects his baseline action with probability $(1 - \epsilon)$ or experiments with a new random action with probability ϵ. That is, for any exploration phase time $t_1 \geq 1$ and for any exploration action time satisfying $m \geq t_2 \geq 1$,
   - $\check{y}_i(t_1, t_2) = y_i^b(t_1)$ with probability $(1 - \epsilon)$,
   - $\check{y}_i(t_1, t_2)$ is chosen randomly (uniformly) over $(Y_i \setminus y_i^b(t_1))$ with probability ϵ.

   Again, the variable ϵ will be referred to as the player's *exploration rate*.
3) **Action Assessment:** After the exploration phase, each player evaluates the average utility received when playing each of his actions during the exploration phase. Let $n_i^{y_i}(t_1)$ be the number of times that player $\mathcal{P}_i$ played action y_i during the exploration phase at time t_1. The average utility for action y_i during the exploration phase at time t_1 if $n_i^{y_i}(t_1) > 0$ is
$$\hat{V}_i^{y_i}(t_1) = \frac{1}{n_i^{y_i}(t_1)} \sum_{t_2=1}^{m} I\{y_i = \check{y}_i(t_1, t_2)\} U_i(\check{y}(t_1, t_2)),$$
otherwise $\hat{V}_i^{y_i}(t_1) = U_{\min}$. The function $I\{\cdot\}$ is the usual indicator function and U_{min} satisfies
$$U_{\min} < \min_i \min_{y \in Y} U_i(y).$$
In words, $U_{\min}$ is less than the smallest payoff any agent can receive.
4) **Evaluation of Better Response Set:** Each player compares the average utility received when playing his baseline action, $\hat{V}_i^{y_i^b(t)}(t_1)$, with the average utility received for each of his other actions, $\hat{V}_i^{y_i}(t_1)$, and finds all played actions which performed δ better than the baseline action. The term δ will be referred to as the players' *tolerance level*. Define $Y_i^*(t_1)$ to be the set of actions that outperformed the baseline action as follows:
$$Y_i^*(t_1) = \left\{ y_i \in Y_i : \hat{V}_i^{y_i}(t_1) \geq \hat{V}_i^{y_i^b(t_1)}(t_1) + \delta \right\}. \quad (2)$$
5) **Baseline Strategy Update:** Each player updates his baseline action as follows:
   - If $Y_i^*(t_1) = \emptyset$, then $y_i^b(t_1 + 1) = y_i^b(t_1)$.
   - If $Y_i^*(t_1) \neq \emptyset$, then
     - With probability ω, set $y_i^b(t_1 + 1) = y_i^b(t_1)$. (We will refer to ω as the player's inertia.)
     - With probability $1 - \omega$, randomly select $y_i^b(t_1 + 1) \in Y_i^*(t_1)$ with uniform probability.
6) Return to Step 2 and repeat.

Before stating the theorem we define the constant $\alpha > 0$ as follows:
$$\alpha := \min\{|U_i(y^1) - U_i(y^2)| > 0 : y^1, y^2 \in Y, \mathcal{P}_i \in \mathcal{P}\}.$$

Theorem 3.3: Let G be a finite n-player weakly acyclic game in which all players use the Sample Experimentation dynamics. For any
- probability $p < 1$,
- tolerance level $\delta \in (0, \alpha)$,
- inertia $\omega \in (0, 1)$, and
- sufficiently small exploration rate $\epsilon > 0$,

if the exploration phase length m is sufficiently large, then for all sufficiently large times $t > 0$, $y(t)$ is a Nash equilibrium of G with at least probability p.

The proof is omitted for brevity. See [1]. Theorem 3.3 can easily be extended to the case where players receive a noisy measurement of their true utility [1], i.e.,
$$\tilde{U}_i(y_i, y_{-i}) = U_i(y_i, y_{-i}) + \nu_i,$$
where ν_i is an i.i.d. random variable with zero mean.

IV. Influencing Nash Equilibria in Resource Allocation Problems

In this section we will derive an approach for influencing the Nash equilibria of a resource allocation problem using the idea of marginal cost pricing. We will illustrate the setup and our approach on a congestion game which is an example of a resource allocation problem.

A. Congestion Game Setup

In order to define a congestion game, we must specify the action set, Y_i, and the utility function, $U_i(\cdot)$, of each player. Towards this end, let R denote a finite set of "resources". For each resource $r \in R$, there is an associated "congestion function" $c_r : \{0, 1, 2, ...\} \to \mathcal{R}$ that reflects the cost of using the resource as a function of the number of players using that resource.

The action set, Y_i, of each player, $\mathcal{P}_i$, is defined as the set of resources available to player $\mathcal{P}_i$, i.e., $Y_i \subset 2^R$, where 2^R denotes the set of subsets of R. Accordingly, an action

$y_i \in Y_i$, reflects a selection of (multiple) resources, $y_i \subset R$. A player is "using" resource r if $r \in y_i$. For an action profile $y \in Y_1 \times \cdots \times Y_n$, let $\sigma_r(y)$ denote the total number of players using resource r, i.e., $|\{i : r \in y_i\}|$. In a congestion game, the utility of player $\mathcal{P}_i$ using resources indicated by y_i depends only on the total number of players using the same resources. More precisely, the utility of player $\mathcal{P}_i$ is defined as

$$U_i(y) = \sum_{r \in y_i} c_r(\sigma_r(y)). \tag{3}$$

Any congestion game with utility functions as in (3) is a potential game [24] with potential function

$$\hat{\phi}(y) = \sum_{r \in R} \sum_{k=1}^{\sigma_r(y)} c_r(k).$$

B. Congestion Game with Tolls Setup

One approach for equilibrium manipulation is to influence players' utilities with tolls [25]. In a congestion game with tolls, a player's utility takes on the form

$$U_i(y) = \sum_{r \in y_i} c_r(\sigma_r(y)) + t_r(\sigma_r(y)),$$

where $t_r(k)$ is the toll imposed on resource r if there are k users.

Suppose that the global planner is interested in minimizing a *general* measure

$$\phi(y) := \sum_{r \in R} f_r(\sigma_r(y)) c_r(\sigma_r(y)), \tag{4}$$

where $f_r : \{0, 1, 2, ...\} \rightarrow \mathcal{R}$ is any arbitrary function. An example of an objective function that fits within this framework is the total congestion experienced by all drivers on the network, which can be evaluated

$$T_c(y) := \sum_{r \in R} \sigma_r(y) c_r(\sigma_r(y)).$$

Proposition 4.1: Consider a congestion game of any network topology. If the imposed tolls are set as

$$t_r(k) = (f_r(k) - 1)c_r(k) - f_r(k-1)c_r(k-1), \quad \forall k \geq 1,$$

then the global planners objective, $\phi(y)$, is a potential function for the congestion game with tolls.

The proof is omitted for brevity. See [1].

V. Illustrative Example – Congestion Game

We will consider a discrete representation of the congestion game setup considered in Braess' Paradox [26]. In our setting, there are 1000 vehicles that need to traverse through the network. The network topology and associated congestion functions are illustrated in Figure 1. Each vehicle can select one of the four possible paths to traverse across the network. The unique Nash equilibrium is when all vehicles take the highlighted route which yields a utility of 2 to each vehicle and a total congestion of 2000.

Since a potential game is weakly acyclic, the payoff based learning dynamics in this paper are applicable learning algorithms for this congestion game. In a congestion game, a payoff based learning algorithms means that drivers have access *only* to the actual congestion experienced. Drivers are unaware of the congestion level on any alternative routes. Figure 1 shows the evolution of drivers on routes when using the Simple Experimentation dynamics with an exploration rate of $\epsilon = 0.25\%$. One can observe that the vehicles' collective behavior does indeed approach that of the Nash equilibrium.

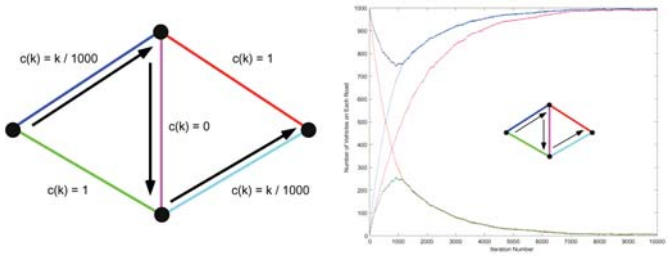

Fig. 1. Illustration of Nash Equilibrium and Evolution of Number of Vehicles on Each Road Using Simple Experimentation Dynamics.

It is easy to verify that this vehicle distribution does not minimize the total congestion experienced by all drivers over the network. The distribution that minimizes the total congestion over the network is when half the vehicles occupy the top two roads and the other half occupy the bottom two roads. The middle road (pink) is irrelevant.

One can employ the tolling scheme developed in the previous section to locally influence vehicle behavior to achieve this objective. In this setting, the new cost functions, i.e. congestion plus tolls, are illustrated in Figure 2, which also shows the evolution of drivers on routes when using the Simple Experimentation dynamics. This simulation used an exploration rate of $\epsilon = 0.25\%$. When using this tolling scheme, the vehicles' collective behavior approaches the new Nash equilibrium which now minimizes the total congestion experienced on the network. The total congestion experienced on the network is now approximately 1500.

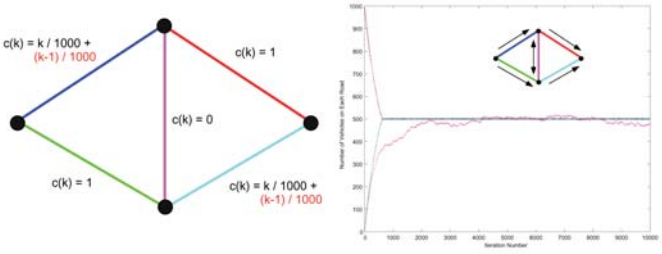

Fig. 2. Illustration of Nash Equilibrium and Evolution of Number of Vehicles on Each Road Using Simple Experimentation Dynamics in Congestion Game with Tolls.

In many applications, players may not have access to their true utility, but do have access to a noisy measurement of their utility. For example, in the traffic setting, this noisy measurement could be the result of accidents or weather conditions. We will revisit the original congestion game (without tolls) as illustrated in Figure 1. We will now assume

that a driver's utility measurement takes on the form

$$\tilde{U}_i(y) = \sum_{r \in y_i} c_r(\sigma_r(y)) + \nu_i,$$

where ν_i is a random variable with zero mean and variance of 0.1. We will assume that the noise is driver specific rather than road specific.

Figure 3 shows a comparison of the evolution of drivers on routes when using the Simple and Sample Experimentation dynamics. The Simple Experimentation dynamics simulation used an exploration rate $\epsilon = 0.25\%$. The Sample Experimentation dynamics simulation used an exploration rate $\epsilon = 0.25\%$, a tolerance level $\delta = 0.002$, an exploration phase length $m = 500000$, and inertia $\omega = 0.85$. As expected, the noisy utility measurements influenced vehicle behavior more in the Simple Experimentation dynamics than the Sample Experimentation dynamics.

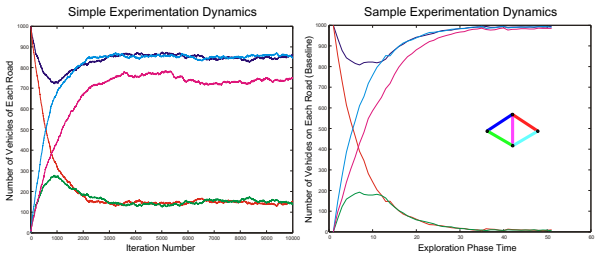

Fig. 3. Comparison of Evolution of Number of Vehicles on Each Road Using Simple Experimentation Dynamics and Sample Experimentation Dynamics (baseline) with Noisy Utility Measurements

VI. Concluding Remarks

We have introduced Safe Experimentation dynamics for identical interest games, Simple Experimentation dynamics for weakly acyclic games with noise-free utility measurements, and Sample Experimentation dynamics for weakly acyclic games with noisy utility measurements. For all three settings, we have shown that for sufficiently large times, the joint action taken by players will constitute a Nash equilibrium. Furthermore, we have shown how to guarantee that a collective objective in a congestion game is a (non-unique) Nash equilibrium.

Our motivation has been that in many engineered systems, the functional forms of utility functions are not available, and so players must adjust their strategies through an adaptive process using only payoff measurements. In the dynamic processes defined here, there is no explicit cooperation or communication between players. On the one hand, this lack of explicit coordination offers an element of robustness to a variety of uncertainties in the strategy adjustment processes. Nonetheless, an interesting future direction would be to investigate to what degree explicit coordination through limited communications could be beneficial.

References

[1] J. R. Marden, H. P. Young, G. Arslan, and J. S. Shamma, "Payoff based dynamics for multi-player weakly acyclic games," *SIAM Journal on Control and Optimization*, 2007, submitted.

[2] A. Ganguli, S. Susca, S. Martinez, F. Bullo, and J. Cortes, "On collective motion in sensor networks: sample problems and distributed algorithms," in *Proceedings of the 44th IEEE Conference on Decision and Control*, Seville, Spain, December 2005, pp. 4239–4244.

[3] V. Borkar and P. Kumar, "Dynamic Cesaro-Wardrop equilibration in networks," *IEEE Transactions on Automatic Control*, vol. **48**, no. 3, pp. 382–396, 2003.

[4] S. B. Gershwin, *Manufacturing Systems Engineering*. Prentice-Hall, 1994.

[5] D. Fudenberg and D. Levine, *The Theory of Learning in Games*. Cambridge, MA: MIT Press, 1998.

[6] S. Hart, "Adaptive heuristics," *Econometrica*, vol. **73**, no. 5, pp. 1401–1430, 2005.

[7] H. P. Young, *Individual Strategy and Social Structure*. Princeton, NJ: Princeton University Press, 1998.

[8] ——, *Strategic Learning and its Limits*. Oxford University Press, 2005.

[9] D. Foster and H. Young, "Regret testing: Learning to play nash equilibrium without knowing you have an opponent," *Theoretical Economics*, vol. **1**, pp. 341–367, 2006.

[10] F. Germano and G. Lugosi, "Global convergence of foster and young's regret testing," *Games and Economic Behavior*, forthcoming.

[11] J. R. Marden, G. Arslan, and J. S. Shamma, "Connections between cooperative control and potential games illustrated on the consensus problem," in *2007 European Control Conference (ECC '07)*, July 2007, to appear.

[12] ——, "Regret based dynamics: Convergence in weakly acyclic games," in *Proceedings of the 2007 International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Honolulu, Hawaii, May 2007, to appear.

[13] R. W. Rosenthal, "A class of games possessing pure-strategy Nash equilibria," *International Journal of Game Theory*, vol. **2**, pp. 65–67, 1973.

[14] J. R. Marden, G. Arslan, and J. S. Shamma, "Joint strategy fictitious play with inertia for potential games," in *Proceedings of the 44th IEEE Conference on Decision and Control*, December 2005, pp. 6692–6697, submitted to *IEEE Transactions on Automatic Control*.

[15] G. Arslan, J. R. Marden, and J. S. Shamma, "Autonomous vehicle-target assignment: A game theoretical formulation," 2006, http://www.seas.ucla.edu/˜shamma.

[16] Y. Shoham, R. Powers, and T. Grenager, "If multi-agent learning is the answer, what is the question?" *Artificial Intelligence*, vol. **171**, no. 7, pp. 365–377, 2007.

[17] S. Mannor and J. Shamma, "Multi-agent learning for engineers," *Artificial Intelligence*, vol. **171**, no. 7, pp. 417–422, 2007.

[18] H. P. Young, "The evolution of conventions," *Econometrica*, vol. 61, no. 1, pp. 57–84, January 1993.

[19] D. Fudenberg and J. Tirole, *Game Theory*. Cambridge, MA: MIT Press, 1991.

[20] D. Monderer and L. Shapley, "Fictitious play property for games with identical interests," *Journal of Economic Theory*, vol. **68**, pp. 258–265, 1996.

[21] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics*. Cambridge, UK: Cambridge University Press, 1998.

[22] J. Weibull, *Evolutionary Game Theory*. Cambridge, MA: MIT Press, 1995.

[23] L. Samuelson, *Evolutionary Games and Equilibrium Selection*. Cambridge, MA: MIT Press, 1997.

[24] D. Monderer and L. Shapley, "Potential games," *Games and Economic Behavior*, vol. **14**, pp. 124–143, 1996.

[25] W. Sandholm, "Evolutionary implementation and congestion pricing," *Review of Economic Studies*, vol. **69**, no. 3, pp. 667–689, 2002.

[26] D. Braess, "Uber ein paradoxen der verkehrsplanning," *Unternehmensforschung*, vol. 12, pp. 258–268, 1968.